Supplemental text for: "Harnessing cloud-computing for biomedical research with Galaxy Cloud"

Cost Analysis Details

When performing the described analysis, we used the Large Instance type available from AWS EC2, which is a virtualized 64-bit computer with 7.5 GB of memory and 2 processing cores clocked at between 2.5 and 2.8 GHz. At the time of writing, the cost of this instance was $0.34 per instance-hour.

Availability.

Complete information on how to use and customize Galaxy CloudMan is available at http://usegalaxy.org/cloud/. The source code for the CloudMan application is open-source and available from http://bitbucket.org/galaxy/cloudman/

Support for customization and extension

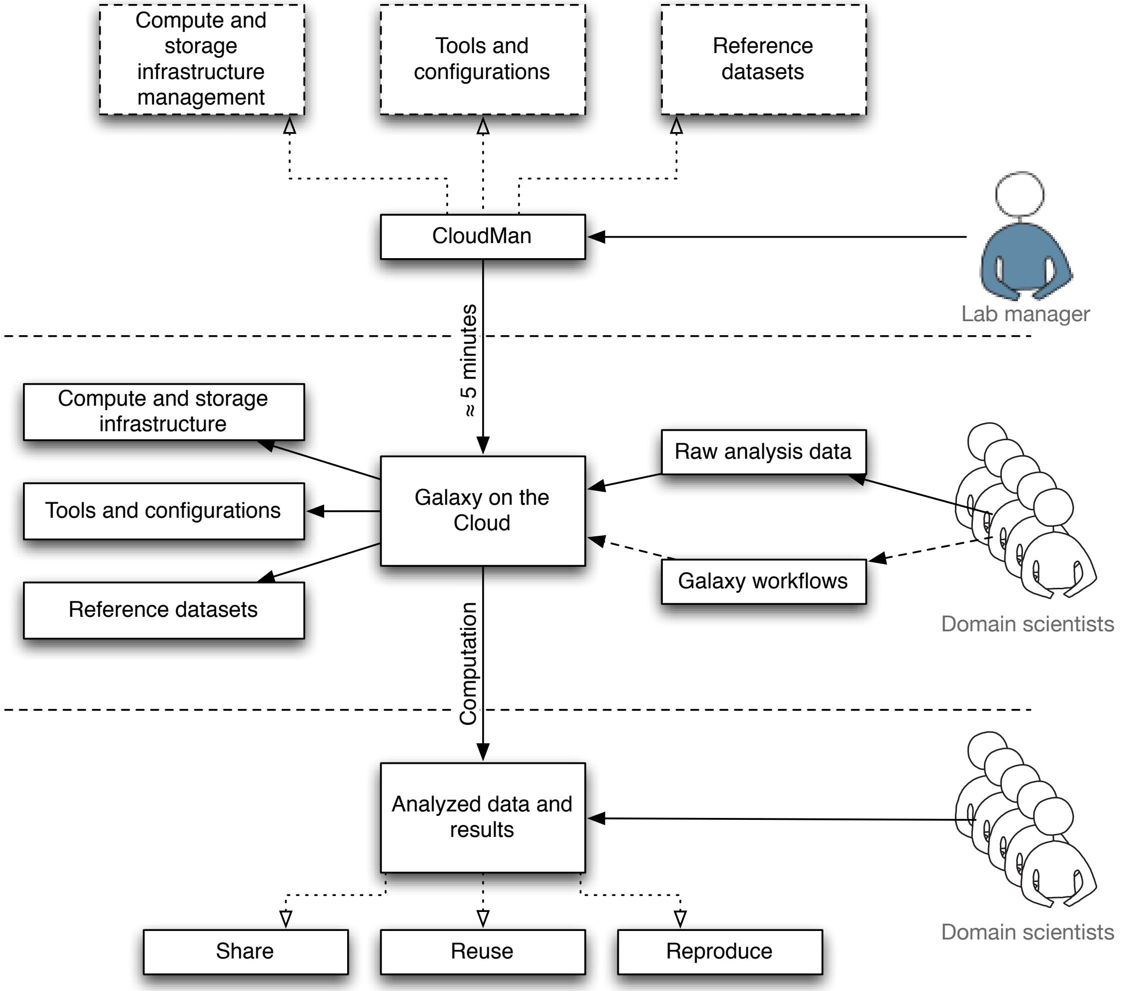
Instances of Galaxy created using Galaxy Cloud can be extended in the same way as any Galaxy instances. Tool and data configuration formats are identical and can be interchanged. Individual cloud instances are entirely self-contained. Once instantiated, the entire cluster configuration is stored in a persistent data repository specific to that cluster instance. Additional or customized tools can be installed on a user-specific volume and persistently enabled through this configuration. Note that this encapsulation facilitates result reproducibility by allowing the exact state including all underlying tools used in an analysis to be saved as a snapshot, as well as allowing for tool versioning. Once instantiated a cluster instance will continue using the same external data resource snapshots, which guarantee reproducibility of analyses because tool versions will remain unchanged. Furthermore, if a configuration update breaks an analysis, it is possible to revert the cluster configuration by utilizing previously working data snapshot.

Appropriateness of Cloud Computing for workloads

Cloud computing has certain deficiencies and limits that merit discussion. In particular, cloud computing may turn out to be considerably more expensive (2 to 3 times) when systems are used 24/7 versus housing and managing systems locally. Additionally, cloud computing offers scalability in terms of infrastructure but not in terms of applications. For example, a 10,000 job workload can be processed quicker by acquiring additional compute resources on demand from the cloud but one job that runs for an entire week will not benefit from the scalability offered through cloud computing. In the context of NGS data, cloud computing may suffer from bandwidth limitations. If input data for a job takes few hours to transfer to the cloud but the job executes only on the order of minutes, the benefit of executing the job on the cloud may be questioned. This problem is alleviated somewhat if the data is deposited in the cloud as it is generated and all analysis is preformed using the cloud provider.

Galaxy CloudMan deployment process.

The flowchart for the high-level CloudMan deployment process that enables composition of infrastructure resources into services required to run Galaxy is diagrammed in Supplementary Figure 1. In summary, a user (typically a lab manager) initiates interaction with the infrastructure through the infrastructure console manager (1) and instantiates a CloudMan machine image (2). CloudMan, as part of the boot process (3), starts and contextualizes itself by obtaining needed context from the persistent data repository (4). The contextualization process involves starting the web server, retrieving the CloudMan application from the persistent data repository and starting it as well as configuring a distributed job manger. Next, based on the information retrieved from the persistent data repository, external storage resources are instantiated, attached to the running instance, and automatically configured (5). This process establishes the necessary infrastructure to run Galaxy, persist user's data, and reduce cost of running a cloud instance. Finally, the Galaxy application is configured with the PostgreSQL database and available reference datasets and started, making it available to users in the same way they would use any other Galaxy instance (6). However, the underlying compute infrastructure managing the workload is automatically scaled to meet the current computational demand (7). As a change in the underlying resources arise, CloudMan handles all aspects of starting/terminating compute instances, exchanging security and file system information, and adding/removing the nodes from the compute cluster. This deployment process and the underlying architecture focuses on clear separation of core components required for a deployment and an instantiation of the respective resources. In addition, the architecture focuses on enabling a completely self-contained solution that does not require an external broker service.



Supplemental Figure 1.